Identifying admixture and genetic ancestry in human populations via genetic drift pattern

Ajai K. Pathak

Estonian Biocentre, Institute of Genomics, University of Tartu, Riia 23b, Tartu 51010, Estonia

*Corresponding author e-mail: ajai.kumar.pathak@ut.ee

ABSTRACT

Progressing technologies have led to the development of numerous methods to analyze the emergent pool of modern and ancient genetic data that decode the pieces of information hidden inside the genome. This can reveal different facets of the demographic history of not just the human populations but other organisms too. However, these methods are so different in their approaches that they can be applied in numerous ways to comprehend the human past at a deeper level. Considering the diversity and versatility in approaches, here I review a set of broadly used methods that follow a specific approach to identify and explain the admixture events, the singular or multiple mixing between two genetically distinct populations. Particularly, I give an overview of the key methods to explicitly detect and quantify admixture by the measure of relative genetic drift observed in populations analyzing genome-wide data, especially focusing on autosomal SNP markers. These methods may not cover the whole picture of human population history; nevertheless, they have significantly transformed our perception about human evolution by unveiling the complexity of the demographic history of human populations to a great deal.

KEYWORDS: Ancient DNA; Archaic hominins; South Asia; Human Evolution; Population Genetics

INTRODUCTION

Archaeological, genetic and linguistic studies so far have provided enough evidence about long-lasting migratory behaviour of the ancestors of modern humans (Gray and Atkinson 2003; Boivin 2007; Reich et al. 2009; Reich et al. 2012; Lazaridis et al. 2014; Haak et al. 2015; Lazaridis et al. 2016; Pagani et al. 2016; Bae et al. 2017). These migrations led the ancient ancestors of modern humans to intermix not just within themselves but even with our distant cousins, the archaic hominins- Neanderthals and Denisovans, after they expanded out of Africa (Green et al. 2010; Reich et al. 2010; Prüfer et al. 2014; Stringer 2016). The period after the advent of farming and pastoralism during the Neolithic period was supposedly the era of multiple migrations and intermixing when we see the pieces of evidence that testify the explode in movements and admixture among different groups of modern humans across most of the global geographical landscapes (Lazaridis et al. 2014; Haak et al. 2015; Lazaridis et al. 2016; Damgaard et al. 2018; Mathieson et al. 2018). Such movements changed the genetic composition of the older people living in different continents by replacing the earlier populations via admixture, consequently shaping the genetic diversity of modern human populations across the world (Reich 2018). In general, admixture between population groups affects the genetic variations and risks of diseases of an admixed population, which receives its genome from the intermixing of distinct populations. Therefore, investigating the ancestral origin of an admixed population offers a prospect to gain insight into the record of ancient forebears of a modern admixed population that may not be surviving at present.

Using ever-evolving statistical tools, we can analyse the growing body of modern and ancient human genomic data to comprehend the demographic history of human populations, consequently extracting the information about the timing and processes involved in shaping the current level of genetic diversities among modern human populations. Mostly, two classes of tools are used frequently; the global-ancestry based methods e.g., PCA (Patterson et al. 2006), STRUCTURE (Pritchard et al. 2000) and ADMIXTURE (Alexander et al. 2009), and local ancestry-based methods e.g., LAMP (Sankararaman et al. 2008), HAPMIX (Price et al. 2009), PCAdmix (Brisbin et al. 2012), fineSTRUCTURE (Lawson et al. 2012). The global ancestry methods identify the population substructure quite robustly, but they cannot be used as formal tests of admixture because of their inability to distinguish if the detected patterns are a result of single admixture history or multiple. Whereas, the local ancestry methods are useful to understand the recent population history, but they are unable to trace the older admixture events of the past (Patterson et al. 2012). Nevertheless, a study by Cavalli-Sfroza and Edwards (Cavalli-Sforza and Edwards 1967) instigated the researchers to develop a group of methods that can model population relationships in addition to formally testing the population admixture histories. These methods analyse the allele frequency patterns among a set of populations and discover population history by comparing the amount of genetic drift in different populations (Reich et al. 2009; Patterson et al. 2012). Here I summarize the basic concepts and applications of a few of key methods that are widely used currently to deduce the admixture history of human populations including the true admixing sources from a list of possible admixing groups, admixture dates, mixture proportions, and relating different ancient populations to a population group of interest via a graph.

Detection and quantification of admixture through f-statistics

Populations with identical allele frequencies i.e., with a common level of genetic variation signify a shared history for such groups, contrary to the populations that are genetically diverged (i.e., populations with different allele frequencies). Exploring such patterns of genetic variation shared

between population groups might provide answers to complex admixture histories of human populations (Pickrell and Reich 2014; Schraiber and Akey 2015). One such prominent tool, f-statistics, first commenced by Reich et al. (Reich et al. 2009) to establish the origin of highly diverged Indian populations by analysing their shared genetic history, has become one of the most widely used methods in modern and ancient DNA studies (Patterson et al. 2012), (Green et al. 2010; Metspalu et al. 2011; Rasmussen et al. 2011; Reich et al. 2012; Allentoft et al. 2015; Haak et al. 2015; Lazaridis et al. 2016; Pathak et al. 2018). The popularity of the f-statistics largely comes from its power to study the admixture events that occurred every so often in the history of human populations (Novembre and Peter 2016). f-statistics that include f2, f3, and f4 statistic, use the allele frequencies observed in a set of two, three and four populations to compute the amount of genetic drift shared between the corresponding number of populations in unrooted tree phylogenies and infer if they share a common population history or not (Patterson et al. 2012; Peter 2016; Harris and DeGiorgio 2017).

Table 1. Drift measuring methods in a nutshell		
Method	Application	Advantages/Limitations
f_3 Admixture	Three population tests to detect if a Test population is derived from the admixture between two populations.	Can detect admixture even when the Test has received equal number of alleles from two admixing groups; But susceptible to the private genetic drift of a population.
f_3 Outgroup	Three population tests to measure the closeness between two populations based on their shared drift.	Unlike other genetic distance measuring approaches like $F_{ST,}$ this method is unaffected by private drift in a population.
f_4-statistics	Validates a suggested tree topology of four populations in unrooted phylogenetic tree.	Detects admixture, but can not specify direction of admixture.
D-statistics	Validates an Outgroup-rooted phylogenetic tree of four populations.	Can detect directional gene flow among four populations.
qpWave	Models a Test population to obtain number of migrations involved in its admixture history.	qpWave and qpAdm may provide confounding results in case of recent gene flow between reference groups and Outgroups.
qpAdm	Estimates the proportions of ancestries a Test has received from distinct migrant groups.	Much robust than earlier used $f_4$ratio to estimate ancestry proportions, especially useful with the ancient genomes.
qpGraph	Relates populations in a Admixture graph and estimate admixture fractions too that best fit to computed f-statistics.	More precise than other graph methods; Requirement of a pre proposed tree limits it to a smaller number of populations.
MixMapper	Extension of qpGraph that builds a graph relating larger populations by assigning admixed and unadmixed groups.	Does not need a pre-proposed tree; Can present migration edges like TreeMix; Much precise than TreeMix; Cannot handle as large data set as TreeMix can.
TreeMix	Iterative Maximum likelihood approach to generate Admixture graph that may contain migrations between population lineages	Can handle larger data set than other graph methods; Confounding estimate of branch length if a population is highly drifted and low-coverage genomes.
ROLLOFF	Dates admixture event from weighted LD curves.	Can date older admixture events than other methods; Slow and prone to sampling bias,

		bottlenecks, and background LD noise.
ALDER	Dates admixture from weighted LD curve and is a statistical test for admixture.	Faster, not prone to background LD noise, sampling bias and bottlenecks; can detect admixture if one of admixing sources is unknown, but dates only recent mixing event in case of continuing admixture in the history of a population.
MALDER	Extension of *ALDER* that detects and dates multiple admixture from distinct sources.	Dates multiple waves of population admixture; cannot distinguish between a gene flow from direct source and that from a group admixed with the direct source.

f-statistics assume that populations sharing identical amounts of genetic drift plausibly share their evolutionary history too (Patterson et al. 2012; Peter 2016; Harris and DeGiorgio 2017). Drift is a change in the occurrence of an allele along different sides (edges) of a graph, thus the drift between two populations is proportional to the difference in observed allelic frequencies of an SNP between those two populations. f2-statistic calculates the amount of genetic drift or genetic difference between two populations like Wright's F_{ST} does. However, f2 is distinct from F_{ST} in two ways; the additive nature i.e., branch edge of phylogenetic tree is the sum of individual branch lengths, and populations fitting in sub-branches farther from the tree root display higher amount of genetic drift relative to the root population since the branch length is higher, but show lower f2 values mutually because the branch length between them is smaller (Patterson et al. 2012; Peter 2016; Harris and DeGiorgio 2017). f2-statistic is useful in applying as the measure of drift that define a graph edge of the admixture graph or a phylogenetic tree.

Testing admixture

f3-statistic also known as f3 admixture test, is a three-population test to evaluate the genetic diversity of a population (A) of interest from two other population groups B and C and assess if population A descends from the mixing between groups B and C. Mostly, if population A derives its ancestry from two groups B and C due to the admixture, the observed allelic frequency of a polymorphism (pA) should remain between allelic frequencies (pB and pC) observed for the same polymorphism in groups B and C, respectively (Patterson et al. 2012; Peter 2016; Harris and DeGiorgio 2017; Wangkumhang and Hellenthal 2018). Thus, statistics f3(A; B, C) is the multiplication of the frequency differences between populations A and B (pA-pB), and A and C (pA-pC), normalized and averaged over all SNPs, so, f3(A; B, C) = (pA-pB) (pA-pC). Since (pA-pB) and (pA-pC) is the amount of genetic drift between populations A and B, and A and C, respectively, the value of statistics f3(A; B, C) is correlated to the comparative drift or branch length between populations A and B, and A and C. In the absence of admixture, the genetic drift along the lineage that produced population A, after its separation from the ancestor population group, must be positive (i.e. > 0). Thus, statistics f3(A; B, C) > 0 implies that population A does not arise out of the admixture between populations B and C. On the other hand, a significant negative value (< 0) of statistics f3(A; B, C) indicates the population A derives from the admixture between populations representing B and C as the allele frequency of population A (pA) falls between that of B (pB) and C (pC). The significance of f3(A; B, C) value is assessed utilizing the weighted-block jack-knife resampling method that first divides the genome into independent chromosomal blocks, then computes the weighted mean value of statistics across all blocks of the same size and

divide it with the standard error of statistics, thus getting a mean Z-score. Nonetheless, we need to be cautious about a false reading of statistics f3(A; B, C) when population A displays a high private drift due to the genetic isolation or bottleneck post admixture event. In such a case, the value of its allele frequency pA will not fall between pB and pC, thus hiding the admixture signal and resulting in a non-negative f3 value (Patterson et al. 2012; Reich et al. 2012; Wangkumhang and Hellenthal 2018).

Measuring genetic relatedness

Raghavan et al. (Raghavan et al. 2014) developed a form of statistics f3, known as 'outgroup-f3', in the form f3 (Outgroup; A, B) to compute the relative genetic attraction between a particular ancient group and a panel of modern populations, finding the contemporary population that is genetically most close to the particular ancient group of interest. In general, the African Yoruba population is considered as the outgroup for all non-Africans (both ancient and modern groups e.g., A and B) and the genetic drift along the lineage of the outgroup is supposed to be constant for all non-Africans due to a long term divergence from the outgroup until the population A and one of the populations from the panel B has some amount of gene flow. Thus, f3(Outgroup; A, B) measures the amount of shared drift of the population A with populations from the given panel B. A higher value of f3(Outgroup; A, B) refers to a higher shared drift between the ancient group (A) and the present-day population from panel B, implying a higher genetic closeness between them. Though outgroup-f3 is also a measure of genetic distance like Wright's F_{ST}, the advantage of the outgroup-f3 is that it is not susceptible to the genetic drift experienced privately by two test populations since their divergence from the outgroup (Raghavan et al. 2014).

Validating a tree topology

Another framework of f-statistics is f4-statistic, a test of treeness that validates a suggested topology of an unrooted phylogenetic tree by computing the amount of shared drift among four populations. If a set of four populations, A and B, and C and D with allele frequencies pA, pB, pC, and pD make a phylogenetic tree, the f4 statistic computes (pA-pB) and (pC-pD), averaged across all loci, to evaluate the correlated variation in the allele frequency differences between populations (A and B) and (C and D), and formally authenticate whether populations (A, B) and (C, D) fit in two distinct branches of the tree.

The three possible unrooted tree topologies: ((A, B), (C,D)), ((A,C), (B,D)), ((A,D), (B,C)) describe four viable relationships for a set of four populations in the absence of any admixture. If topology ((A,B), (C,D)) is correct, the value of statistics f4(A,B;C,D) should be 0 because when there is no admixture, the difference in the allele frequencies between populations A and B i.e., (pA-pB) is unaffected by allele frequency difference between C and D i.e., (pC-pD), implying the genetic drift between populations (A and B) and (C and D) are not linked. Thus, in the context of all viable tree topologies, if the value of f4-statistic deviates significantly from 0, it indicates a possible admixture scenario between populations of two branches consequently, a simple phylogenetic tree connecting these four populations is impossible if we rule out admixture. A significant positive value of f4-statistic in the form f4 (A,B;C,D) suggest a higher genetic affinity between populations A and C or B and D; however, an additional closeness in genetic ancestry between populations B and C or A and D gives a significant but negative value of f4-statistic. In one way, f4-statistics is capable of testing the admixture as the f3-statistic does, but it cannot specify the direction of admixture that f3-statistic can.

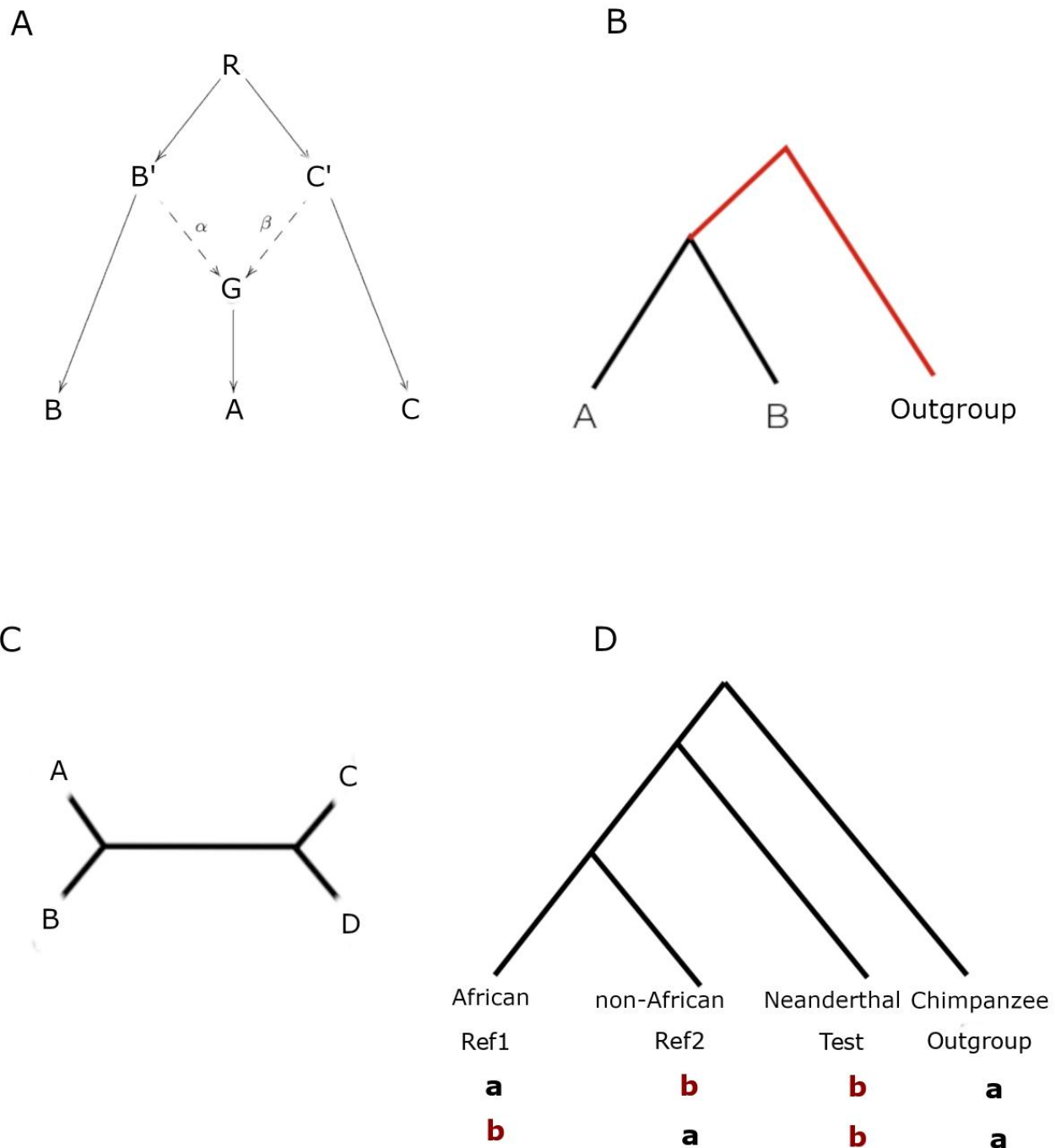Figure 1. Representation of f- and D-statistics: (A) phylogenetic tree of simple admixture where B' and C' split from a root population R. Population A's ancestral lineage G was formed by admixture (between B' and C') in proportions α: β, (β = 1−α). Contemporary populations A, B, C are formed by drift from their ancestors (G, B' and C' respectively), thus f3(A; B, C) will give negative value; (B) Outgroup statistic f3(Outgroup; A, B) measures the closeness between populations A and B, as measured from outgroup, the red colour indicates the branch length from outgroup to the common ancestor of A and B: Higher value of the statistic indicates more genetic similarity between A and B; (C) Unrooted four-population tree topology showing the relationships between populations A, B, C, and D, wherein A, B and C, D are

forming two individual clades. The result f4(A, B; C, D) = 0 indicate that the tree is consistent, while a value beyond 0 indicates otherwise; (D) D-statistic representing a rooted four-population asymmetric tree relationship between modern humans (Africans or, non-Africans) and Archaic hominins (Neanderthal) using Chimpanzee as the outgroup, the same can be applied for any four population combinations (Ref1, Ref2, Test, Outgroup) in Patterson's D-statistic. abba configuration shows derived allele sharing between Ref2 and Test that will increase the value of D-statistics, baba configuration indicates derived allele sharing between Ref1 and Test decreasing the value of D-statistic. Adapted from Green et al. 2010 and Patterson et al. 2012.

Also, consider the admixture scenario when population D received identical amount of genetic ancestry from two populations (B and C), while the third population (A) is an outgroup; in such a situation, f4-statistics might not be able to detect the admixture, but f3-statistics truly infers the occurrence of admixture (Patterson et al. 2012; Reich et al. 2012; Harris and DeGiorgio 2017). An application of f4-statistic is f4 ratio (Reich et al. 2009; Patterson et al. 2012; Reich et al. 2012) that estimates the amount of ancestry contributed by a probable ancestral group to the admixed population of interest. In f4 ratio, we use five population groups (A, B, C, D, O) to first obtain two sets of f4-statistic values, combining four populations in each set, and then divide them in a form p = f4(B,O;A,D)/ f4 (B,O;C,D), where p denotes the value of f4 ratio, suggesting the amount of ancestry the population A has received from the ancestral population group C (Reich et al. 2009; Harris and DeGiorgio 2017).

Number of migrations and ancestry proportions

Lately developed methods qpWave and qpAdm apply the common ideas associated with f4-statistics (Reich et al. 2012; Moorjani et al. 2013; Lazaridis et al. 2014; Haak et al. 2015). In qpWave, we first learn the least number of reference groups that are possibly involved in admixture history of a certain admixed test population, and later use qpAdm to estimate the proportions of ancestries the admixed test population has obtained from those reference groups. qpWave is also known as f4 rank test that calculates the rank for the matrix of f4 statistics with dimension p x q, where p

indicates a list of possible reference populations to be tested and q is the number of suggested outgroups without any back-gene flow from the reference populations. f4 rank of the matrix will be 0 when a population being tested has no admixture history, but each added mixing event in the shared population history of a set of test and reference populations will raise f4 rank of matrix by 1. Thus, qpWave gets the number of migrations affecting the admixture history of a target Test population. The method assumes the set of reference populations are related to the purported outgroup populations with a differential amount of shared genetic drift and no gene flow occurred between the suggested outgroups and either the test or reference populations (Reich et al. 2012; Moorjani et al. 2013; Lazaridis et al. 2014; Haak et al. 2015). qpAdm applies f4-statistics in a regression context to obtain the admixture coefficients for a test population by combining methods from qpWave. qpAdm framework first assumes that a test population received proportions α1, α2,......αn from a set of reference populations Xa, Xb,...Xn, and calculates the ancestry proportions each mixing reference group passes to the test population, examining the varying degree of genetic relationships (drift) between reference populations and a fitting number of outgroups without explicit modelling. Both the qpWave and qpAdm assume a deep evolutionary history between reference populations and outgroups without any recent gene flow between suggested set of outgroups and either the test or reference population. Therefore, caution must be taken while choosing reference and outgroup populations. qpAdm

framework has become a popular method, especially in the context of flourishing aDNA studies, explaining the ancient inherited elements of extant continental populations (Reich et al. 2012; Haak et al. 2015; Lazaridis et al. 2016; Damgaard et al. 2018; Narasimhan et al. 2019).

Spotting gene flow via D-statistics

Green et al. (Green et al. 2010) presented another four-population approach to detect the introgression of Neanderthal ancestry in modern humans, known as D-statistics. It is a tool that validates an outgroup-rooted phylogenetic tree of four populations, where other three populations belong to the same in-group taxa and recognize the interbreeding between a non-sister lineage population (Neanderthal) and either of two sister lineage populations (Africans, non-Africans) by the measure of shared derived alleles. Since derived alleles are mutations that accumulate in time after the split of the lineages of in-group taxa from the outgroup, both the sister lineage populations should share an equal number of derived alleles with the third group of non-sister lineage until there is some introgression. Higher derived allele sharing between the non-sister group and either of the sister groups indicates admixture. Thus, 0 value of D-statistics in form D(Africans, non-Africans, Neanderthal, Chimpanzee) indicates no Neanderthal related introgression in either group of modern humans. While in the case of hybridization between Neanderthal and either Africans or non-African sister groups, the value of D-statistics will deviate from 0. Green's D-statistics was a simple model of admixture that occurred between a fixed number of people related to ancient ancestors of the populations to be tested, assuming random mating amongst them, but inferring admixture could be alarming when biased mixing happened between populations of different sizes. However, an extended version of D-statistics introduced later by Durand et al. (Durand et al. 2011) and Patterson et al. (Patterson et al. 2012) assumed the model of admixture between different number of individuals related to

non-randomly mating ancestral groups and in turn solved the risk of wrong inference if ancient groups themselves show a structured pattern. This extended D-statistics comes with a few essential advantages; It can analyse genotype data because of its robustness to the ascertainment bias unlike the genome sequence necessity applied to Green's D-statistics, and also the Patterson's D-statistics is applicable to detect admixture for any combination of four populations not just the introgression from archaic specimens. D-statistics is also called as abba/baba approach, indicating a sharing pattern of ancestral (a) and derived (b) alleles along the branches of a purported phylogenetic tree, where $D = Nabba - Nbaba / Nabba + Nbaba$; Nabba and Nbaba are the number of abba and baba sites across the genome. If we consider a D-statistic of form D (Ref1, Ref2, Test, Outgroup), D = 0 when abba and baba sites are equal in number; higher number of abba sites, showing higher derived allele sharing between Ref2 and admixing test group, would raise the value of D beyond 0 (positive), while higher derived allele sharing between Ref1 and admixing test population, stated by higher baba sites, would decrease the value of D (negative) (Patterson et al. 2012; Harris and DeGiorgio 2017). The value of D-statistics lies in the range of -1 to +1 unlike the f4-statistic, and the significance of both D and f-statistics is conditional to the Z-score (> 3 or < -3), calculated by weighted block jack-knife approach (discussed earlier).

Population splits and admixture via graph construction approaches

The above-mentioned statistical tools reflect the power of statistics that compute the amount of genetic drift and explain how different populations are related to each other. Applying these statistical approaches, we may also construct various graph-like representations that relate population splits and gene flow for a bigger set of populations than f and D-statistics alone can handle. These graph building approaches indeed are complementary to

the results obtained from f- and D-statistics that analyse the genetic diversities of many populations and suggest an elaborate tree-like topology, illustrating their mutual relationships. Graph-based techniques analyse multiple population-level genomes to obtain the allele frequencies and build a complete tree of populations based on drift parameters which explain the complex evolutionary histories of these populations linking different episodic migrations and admixtures (Pickrell and Pritchard 2012; Lipson et al. 2013; Harris and DeGiorgio 2017; Wangkumhang and Hellenthal 2018). I will discuss three widely used tools, TreeMix (Pickrell and Pritchard 2012), qpGraph (Patterson et al. 2012), and MixMapper (Lipson et al. 2013). TreeMix applies the maximum likelihood approach to build a graph with directed networks of multiple populations underlying the best fitting network of branch lengths. These branch lengths compare the degree of genetic drifts (history) among a group of populations (Pickrell and Pritchard 2012). Supposing the evolution without selection (i.e., neutral), TreeMix uses multivariate normal distribution on allele frequency data (for both the biallelic loci and microsatellite data) to model the relative occurrence of different allelic variants observed among a group of populations. In general, descendant populations contain a similar amount of mean allele frequencies as their ancestor, and a difference in allele frequencies of two given populations is associated with the difference in their shared genetic drift relative to their ancestor. In the iterative approach of TreeMix, first a maximum likelihood tree is selected out of possible trees formed from three randomly picked populations during each repetition, to which residual groups are added randomly one after the other. Finally, the tree is locally rearranged to evaluate the fit to the data and the best tree that rises the likelihood is considered (Pickrell and Pritchard 2012; Harris and DeGiorgio 2017). The automated fitting procedure enables TreeMix to appropriately handle larger trees with several possible admixture scenarios at the same time; for

example, a graph containing mixture events defined by the user or without any migration event. Though TreeMix graphs can easily deduce the degree of population differentiation in case of good quality modern genomes, it is not that straight with ancient specimens especially with the low-quality genomes. Thus, it is recommended to avoid construing branch lengths in such cases; however, the positioning of populations on the tree remains useful even with the ancient genomes (Raghavan et al. 2014).

In the case of other tree-building tools like qpGraph (Patterson et al. 2012) and its generalized version MixMapper (Lipson et al. 2013), users need to define the admixed and unadmixed population groups beforehand; therefore, they are better suited to apply on a set of specified populations. qpGraph analyses the genotype data and a suggested tree relating all populations (including mixed and non-mixed) and provides estimated branch lengths and admixture fractions that best fit the calculated allele frequency statistic (f-statistics). On the contrary, MixMapper fits populations in two phases; initially, it builds a neighbour-joining tree scaffold of unadmixed groups, identified by f3-statistic, and then adds admixed groups on to this early tree scaffold based on best-fitting parameter values. MixMapper does not need a hypothesized tree relating populations in advance as is the case with qpGraph, making it more suitable to infer ancestry in larger data sets like TreeMix (Lipson et al. 2013; Harris and DeGiorgio 2017). Similar to TreeMix, MixMapper is also capable of identifying three-way admixtures and presenting gene flow (or migration) between population lines. Although the admixture graph produced by MixMapper might be more precise than that of TreeMix due to the additional ability of MixMapper in assigning admixed and unadmixed populations; nevertheless, this ability also restricts MixMapper from analysing a data set as large as TreeMix can (Harris and DeGiorgio 2017). Seeing differential suitability of each graph constructing methods to different contexts, selection of the best method to

produce a graph for a given set of populations mostly relies on prior knowledge about the complexity of population histories of target groups (Lipson et al. 2013).
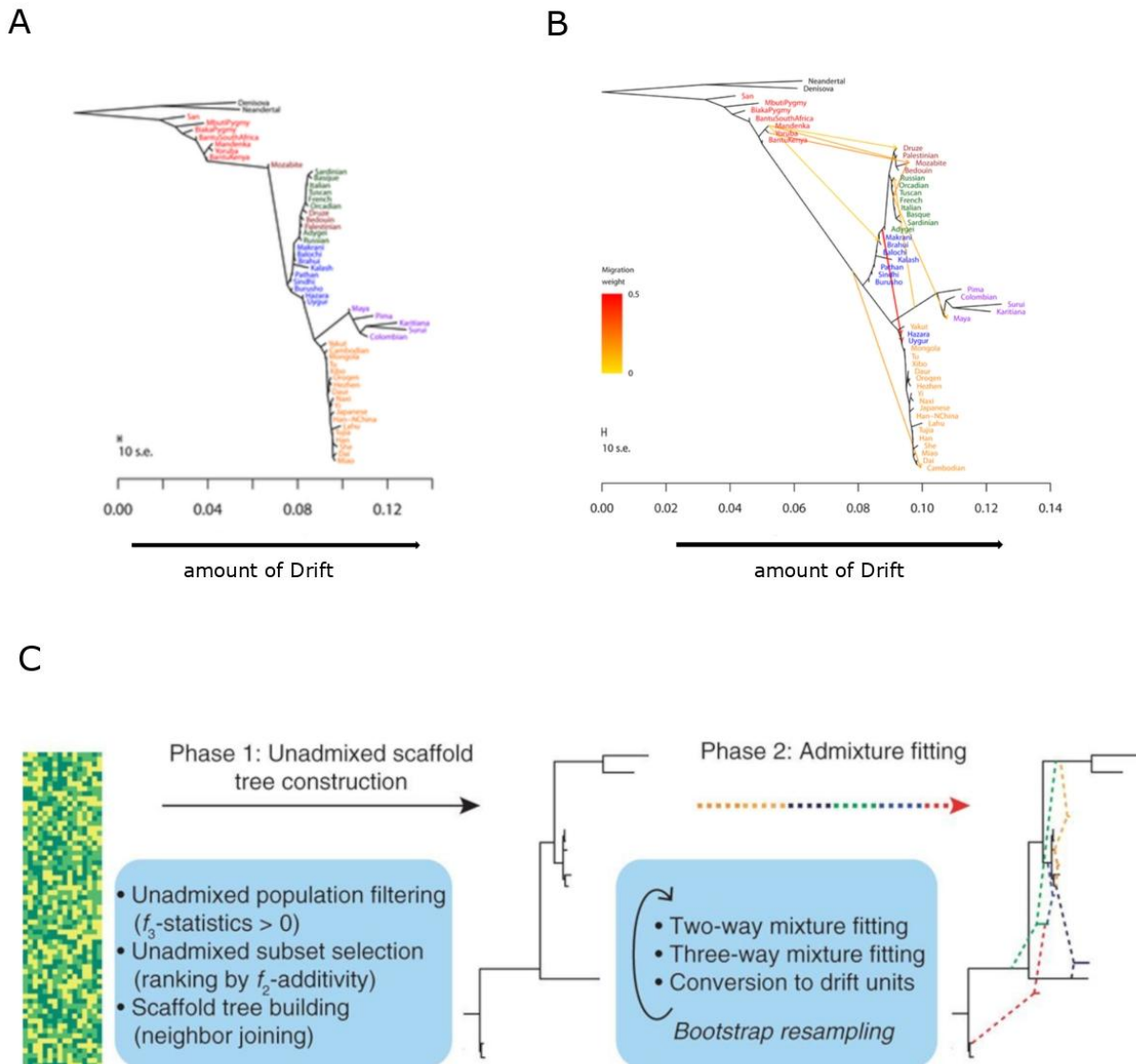

Figure 2. Inferred human phylogenetic tree fitting admixture events. The structure of TreeMix and MixMapper constructed graph plots taken, respectively, from Pickrell et al. (2012) and Lipson et al. (2013). TreeMix plots comprise archaic hominins and modern human population groups: (A) maximum likelihood tree relating all modern humans and archaic hominins without considering any migration (B) tree allowing for ten migration events that occurred between different continental groups of modern humans. The colour of migration edges reflects the migration weight while horizontal branch lengths correlate to the corresponding degree of genetic drifts each branch has faced; Graph fitting mixture parameters by MixMapper: (C) this two-phase method initially builds a tree of unadmixed populations and then tries to fit the remaining populations as admixtures. In the first phase, MixMapper yields a ranking of likely unadmixed trees in order of deviation from f2-additivity, then this list is used to select a tree as a scaffold. In the final phase, MixMapper attempts fitting remaining populations as two- or three-way mixtures

between branches of the unadmixed tree. In each case, MixMapper applies bootstrap resampling to obtain collective predictions, thus enabling confidence estimation for inferred results.

Tracking the time of admixture

Applying the aforementioned approaches, we may answer a range of questions related to the impact of past migrations and mixings for a population of interest, but a comprehensive sketch of the admixture history associated with that population will not be complete without knowing the time of occurrence of the admixture event. I briefly discuss here, ROLLOFF (Moorjani et al. 2011; Patterson et al. 2012), ALDER (Loh et al. 2013) and MALDER (Pickrell et al. 2014), the widely used approaches for admixture dating. They exploit the drift and linkage disequilibrium (LD) obtained from the allele frequencies of a set of populations to infer the date of the admixture, a population of interest has encountered. Linkage disequilibrium (LD) is the non-random pattern of association among alleles at more distant loci. Recombination events that occur at least once in a sufficiently longer genome split up specific alleles from one another, thus the admixture created LD between allelic sites that are farther away from each other decays each generation. Thus, the size of genome segments with a haplotype that contains two neutrally evolving markers decline as a function of time since admixture, which can be applied to get the admixture date. Both ROLLOFF and ALDER model the exponential LD decay among pairs of SNPs on the same chromosome compared to their genetic distances, weighted by the ability of each marker to differentiate between proxies of the admixing source groups, which is called as weighted LD statistic (Moorjani et al. 2011; Loh et al. 2013; Harris and DeGiorgio 2017; Wangkumhang and Hellenthal 2018).

Owing to a new algorithm implementation in ALDER, it is not just faster than ROLLOFF, but also immune to sampling bias and bottlenecks or recent admixture events in populations. An additional advantage of the ALDER over ROLLOFF comes from its capability to calculate the admixture date from a single reference and the admixed population in case one of the admixing groups is unknown. ALDER also nullifies the obscure inference due to the background LD noise by fitting the weighted LD curve for sufficiently distanced pairs of loci. ALDER and ROLLOFF, respectively, can date the admixture events of 300 and 500 generations old; however, they assume a single pulse of admixture and thus a continuous mixing pattern in populations history of a group might reveal the most recent admixture event (Patterson et al. 2012; Loh et al. 2013; Harris and DeGiorgio 2017). Lately, a method MALDER (Pickrell et al. 2014) extending the ideas of ALDER (Loh et al. 2013) was developed to date the multiple admixture events that involved different admixing sources. Considering a variety of reference admixing sources, MALDER uses a mixture of exponential LD decay curves instead of the single exponential decay curve as a function of genetic distance; therefore, formally testing the existence and times of multiple mixing events.

Summarizing the global population histories through drift and conclusions

Applying the earlier discussed approaches to countless genomes from modern and ancient human populations enabled studies to unravel distinct ancient admixture and migration patterns that shaped the population dynamics across the world. These tools advanced the investigations of human population history to a greater extent, providing the genetic evidence of the interbreeding between the ancestors of modern non-African humans and archaic hominins (Neanderthal and Denisovans), revealing the ancient population structures in West Eurasia and complex demographic histories of human populations across Eurasia, apart from characterizing the genetic component of ancient Indus Valley civilization people in South Asia.

Drift based approaches, especially D-statistics, played a pivotal role in discovering that almost all modern non-Africans harbour a small portion of ancestry from Neanderthal owing to the interbreeding events between their ancestors and Neanderthal that occurred outside of Africa around 50–60 KYA (Green et al. 2010; Sankararaman et al. 2012; Prüfer et al. 2014; Vernot and Akey 2015; Skoglund and Mathieson 2018). Later, the evidence of introgression related to other archaic hominins, Denisovans, was observed in modern humans living in Melanesia, and to mainland East and South Asia as well, but at a lower level (Reich et al. 2010; Skoglund and Jakobsson 2011; Meyer et al. 2012; Qin and Stoneking 2015; Browning et al. 2018).

However, it seems that the ancient ancestors of modern humans were great voyagers which resulted in continuous population movements to different territories and episodes of admixture with each other as well, making the demographic histories of continental populations quite complex and interesting. Combining different genetic drift measuring tools to analyse the genomes of modern and ancient individuals have revealed great details of the present and past population histories of America, Africa and Eurasia (Reich et al. 2009; Tishkoff et al. 2009; Metspalu et al. 2011; Pagani et al. 2012; Moorjani et al. 2013; Lazaridis et al. 2014; Pickrell et al. 2014; Raghavan et al. 2014; Skoglund et al. 2014; Allentoft et al. 2015; Haak et al. 2015; Mathieson et al. 2015; Raghavan et al. 2015; Basu et al. 2016; Lazaridis et al. 2016; Skoglund et al. 2016; Chaubey et al. 2017; Damgaard et al. 2018; Lazaridis 2018; Moreno-Mayar et al. 2018; Pathak et al. 2018; Narasimhan et al. 2019; Shinde et al. 2019).

Interestingly, in the case of South Asia that harbours the second highest genetic diversity after Africa, these approaches have been greatly helpful to augment the knowledge regarding the complex demographic history of modern South Asian populations. Some of the f-statistics framework tools were initially developed to explain the

populations histories of Indian people, resulting into the claim that all contemporary Indian populations mainly comprise of two genetic components, the West Eurasian related ANI (Ancient North Indian) and the autochthonous ASI (Ancestral South Indian) (Reich et al. 2009). Both the ancestral ANI and ASI groups were supposedly admixed around 1.9-4.2 KYA ago (Moorjani et al. 2013), deriving modern South Asians who later shifted to endogamy practices (Moorjani et al. 2013; Basu et al. 2016). These tools traced the approximate date of arrival of several historical migrants to India (Chaubey et al. 2016; Chaubey et al. 2017). They also detected the admixture event at 2-3.8 KYA between ancestors of Indian Austroasiatic and Southeast Asians that genetically shaped the modern Munda populations of India, besides finding the closest proxy of Southeast Asian surrogate using qpAdm (Tätte et al. 2019). Applying such approaches on modern genomes, a few studies successfully modelled the current South Asian populations using ancient individuals from West Eurasia and the first set of ancient DNA from South Asia (Pathak et al. 2018; Narasimhan et al. 2019; Pathak et al. 2019; Shinde et al. 2019). One such study (Pathak et al. 2018) also detected the distinctively high genetic affinity of the historical Indian group, Ror, with the Bronze Age Steppe people and modern Europeans. In addition, they also observed that affinity of Ror to Europeans is correlated to the degree of Steppe component in Europeans. Combining drift measuring tools to analyse the first aDNAs from South Asia have served well to untangle the hidden facts about the prehistorical peopling of South Asia (Narasimhan et al. 2019; Shinde et al. 2019). Shinde et al. (Shinde et al. 2019) inferred that modern South Asians are largely the descendants of the Bronze Age Indus people; however, the Steppe ancestry that is prevalent in contemporary South Asians, especially in North and Northwestern Indians, lacked in the Bronze Age IVC people. They claimed that IVC people were comprised of genetic components from Southeast Asian hunter-gatherers (Onge like) and ancient Iran related ancestry that was present

before the split between the ancient Iranian farmers and hunter-gatherers i.e., before the rise of farming in Fertile Crescent; thus, postulating that the development of agriculture in South Asia happened either indigenously or through the exchange of ideas. While the other study (Narasimhan et al. 2019) observed a complicated pattern of migrations and mixings between Central and South Asian groups during the prehistorical era, and one such migration led admixture during the Late and Middle Bronze age is mostly responsible for the spread of Steppe ancestry to South Asia.

Contamination in genomes might mimic the admixture signal, resulting into the inaccurate inference of complex population histories of humans, especially with the low-quality ancient genomes. Most of current methods might not be able to differentiate between a true signal of admixture and artefacts; therefore, developing methods that ignore contamination bias will be helpful. The aforementioned outgroup f3-statisitic has been the key to address such problems, it is quite robust to contamination, providing the precise genetic similarity between populations. Also, presently available admixture graph methods have some limitations, e.g., the inability of qpGraph to infer the tree topologies and graph parameters simultaneously, while TreeMix and MixMapper need to examine a delimited space of potential admixtures. Therefore, it will be helpful to develop a robust, fast and automated admixture graph tool that can deduce the phylogenetic tree and graph parameters, in parallel, for a larger sample size. However, the approaches to measure genetic drifts and related advances will stay crucial in faster and more accurate analyses of a growing number of human genomes, both modern and ancient, providing the essential analytical strength to our hunt of a greater and deeper understanding of the unresolved factors that shaped the rich population histories of humans across the world. Particularly, the cumulative genomes of ancient individuals from the vast tropical regions that remained unexplored until late due to the environmental constraints, now showing the potential in extraction of ancient genomes from such regions owing to the advancing technologies, will prove beneficial in combination with drift measuring approaches and provide a more complete picture of human movements and admixtures. Additionally, given the accessibility to new genome and phenotype data from both the modern humans and ancient hominins, it will be essential to develop new methodologies that can jointly model both the admixture and selection.

Acknowledgements

Conflict of Interest

The author declares that no competing interest exists.

REFERENCES

Alexander DH, Novembre J, Lange K. 2009. Fast model-based estimation of ancestry in unrelated individuals. Genome Res [Internet] 19. Available from: https://doi.org/10.1101/gr.094052.109

Allentoft ME, Sikora M, Sjögren K-G, Rasmussen S, Rasmussen M, Stenderup J, Damgaard PB, Schroeder H, Ahlström T, Vinner L. 2015. Population genomics of bronze age Eurasia. Nature 522:167–172.

Bae CJ, Douka K, Petraglia MD. 2017. On the origin of modern humans: Asian perspectives. Science 358:eaai9067.

Basu A, Sarkar-Roy N, Majumder PP. 2016. Genomic reconstruction of the history of extant populations of India reveals five distinct ancestral components and a complex structure. Proc Natl Acad Sci U S A [Internet] 113. Available from: https://doi.org/10.1073/pnas.1513197113

Boivin N. 2007. Anthropological, historical, archaeological and genetic perspectives on the origins of caste in South Asia. In: The Evolution and History of Human Populations in South Asia. Springer. p. 341–361.

Brisbin A, Bryc K, Byrnes J, Zakharia F, Omberg L, Degenhardt J, Reynolds A, Ostrer H, Mezey JG, Bustamante CD. 2012. PCAdmix: Principal Components-Based Assignment of Ancestry Along Each Chromosome

in Individuals with Admixed Ancestry from Two or More Populations. hbio 84:343–364.

Browning SR, Browning BL, Zhou Y, Tucci S, Akey JM. 2018. Analysis of Human Sequence Data Reveals Two Pulses of Archaic Denisovan Admixture. Cell 173:53-61.e9.

Cavalli-Sforza LL, Edwards AWF. 1967. Phylogenetic Analysis: Models and Estimation Procedures. Evolution 21:550–570.

Chaubey G, Ayub Q, Rai N, Prakash S, Mushrif-Tripathy V, Mezzavilla M, Pathak AK, Tamang R, Firasat S, Reidla M, et al. 2017. "Like sugar in milk": reconstructing the genetic history of the Parsi population. Genome Biology 18:110.

Chaubey G, Singh M, Rai N, Kariappa M, Singh K, Singh A, Pratap Singh D, Tamang R, Selvi Rani D, Reddy AG, et al. 2016. Genetic affinities of the Jewish populations of India. Scientific Reports 6:19166.

Damgaard P de B, Martiniano R, Kamm J, Moreno-Mayar JV, Kroonen G, Peyrot M, Barjamovic G, Rasmussen S, Zacho C, Baimukhanov N, et al. 2018. The first horse herders and the impact of early Bronze Age steppe expansions into Asia. Science:eaar7711.

Durand EY, Patterson N, Reich D, Slatkin M. 2011. Testing for Ancient Admixture between Closely Related Populations. Mol Biol Evol 28:2239–2252.

Gray RD, Atkinson QD. 2003. Language-tree divergence times support the Anatolian theory of Indo-European origin. Nature 426:435–439.

Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, Kircher M, Patterson N, Li H, Zhai W, Fritz MH-Y, et al. 2010. A Draft Sequence of the Neandertal Genome. Science 328:710–722.

Haak W, Lazaridis I, Patterson N, Rohland N, Mallick S, Llamas B, Brandt G, Nordenfelt S, Harney E, Stewardson K. 2015. Massive migration from the steppe was a source for Indo-European languages in Europe. Nature 522:207–211.

Harris AM, DeGiorgio M. 2017. Admixture and Ancestry Inference from Ancient and Modern Samples through Measures of Population Genetic Drift. Hum. Biol. 89:21–46.

Lawson DJ, Hellenthal G, Myers S, Falush D. 2012. Inference of population structure using dense haplotype data. PLoS genetics 8:e1002453.

Lazaridis I. 2018. The evolutionary history of human populations in Europe. Current Opinion in Genetics & Development 53:21–27.

Lazaridis I, Nadel D, Rollefson G, Merrett DC, Rohland N, Mallick S, Fernandes D, Novak M, Gamarra B, Sirak K. 2016. Genomic insights into the origin of farming in the ancient Near East. Nature 536:419–424.

Lazaridis I, Patterson N, Mittnik A, Renaud G, Mallick S, Kirsanow K, Sudmant PH, Schraiber JG, Castellano S, Lipson M, et al. 2014. Ancient human genomes suggest three ancestral populations for present-day Europeans. Nature 513:409.

Lipson M, Loh P-R, Levin A, Reich D, Patterson N, Berger B. 2013. Efficient Moment-Based Inference of Admixture Parameters and Sources of Gene Flow. Mol Biol Evol 30:1788–1802.

Loh P-R, Lipson M, Patterson N, Moorjani P, Pickrell JK, Reich D, Berger B. 2013. Inferring Admixture Histories of Human Populations Using Linkage Disequilibrium. Genetics 193:1233–1254.

Mathieson I, Alpaslan-Roodenberg S, Posth C, Szécsényi-Nagy A, Rohland N, Mallick S, Olalde I, Broomandkhoshbacht N, Candilio F, Cheronet O, et al. 2018. The genomic history of southeastern Europe. Nature 555:197–203.

Mathieson I, Lazaridis I, Rohland N, Mallick S, Patterson N, Roodenberg SA, Harney E, Stewardson K, Fernandes D, Novak M. 2015. Genome-wide patterns of selection in 230 ancient Eurasians. Nature 528:499–503.

Metspalu M, Romero IG, Yunusbayev B, Chaubey G, Mallick CB, Hudjashov G, Nelis M, Mägi R, Metspalu E, Remm M. 2011. Shared and unique components of human population structure and genome-wide signals of positive selection in South Asia. The American Journal of Human Genetics 89:731–744.

Meyer M, Kircher M, Gansauge M-T, Li H, Racimo F, Mallick S, Schraiber JG, Jay F, Prüfer K, Filippo C de, et al. 2012. A High-Coverage Genome Sequence from an Archaic Denisovan Individual. Science 338:222–226.

Moorjani P, Patterson N, Hirschhorn JN, Keinan A, Hao L, Atzmon G, Burns E, Ostrer H, Price AL, Reich D. 2011. The History of African Gene Flow into Southern Europeans, Levantines, and Jews. PLOS Genetics 7:e1001373.

Moorjani P, Thangaraj K, Patterson N, Lipson M, Loh PR, Govindaraj P, Berger B, Reich D, Singh L. 2013. Genetic evidence for recent population mixture in India. Am J Hum Genet [Internet] 93. Available from: https://doi.org/10.1016/j.ajhg.2013.07.006

Moreno-Mayar JV, Potter BA, Vinner L, Steinrücken M, Rasmussen S, Terhorst J, Kamm JA, Albrechtsen A, Malaspinas A-S, Sikora M, et al. 2018. Terminal Pleistocene Alaskan genome reveals first founding population of Native Americans. Nature 553:203–207.

Narasimhan VM, Patterson N, Moorjani P, Rohland N, Bernardos R, Mallick S, Lazaridis I, Nakatsuka N, Olalde I, Lipson M, et al. 2019. The formation of human populations in South and Central Asia. Science 365.

Novembre J, Peter BM. 2016. Recent advances in the study of fine-scale population structure in humans. Current Opinion in Genetics & Development 41:98–105.

Pagani L, Kivisild T, Tarekegn A, Ekong R, Plaster C, Gallego Romero I, Ayub Q, Mehdi SQ, Thomas MG, Luiselli D, et al. 2012. Ethiopian genetic diversity reveals linguistic

stratification and complex influences on the Ethiopian gene pool. Am. J. Hum. Genet. 91:83–96.

Pagani L, Lawson DJ, Jagoda E, Mörseburg A, Eriksson A, Mitt M, Clemente F, Hudjashov G, DeGiorgio M, Saag L. 2016. Genomic analyses inform on migration events during the peopling of Eurasia. Nature 538:238–242.

Pathak AK, Kadian A, Kushniarevich A, Montinaro F, Mondal M, Ongaro L, Singh M, Kumar P, Rai N, Parik J, et al. 2018. The Genetic Ancestry of Modern Indus Valley Populations from Northwest India. The American Journal of Human Genetics 103:918–929.

Pathak AK, Srivastava A, Singh PP, Das D, Bandopadhyay A, Singh P, Chaubey G. 2019. Historic migration to South Asia in the last two millennia: A case of Jewish and Parsi populations. J Biosci 44:72.

Patterson N, Moorjani P, Luo Y, Mallick S, Rohland N, Zhan Y, Genschoreck T, Webster T, Reich D. 2012. Ancient admixture in human history. Genetics 192:1065–1093.

Patterson N, Price AL, Reich D. 2006. Population structure and eigenanalysis. PLoS genet 2:e190.

Peter BM. 2016. Admixture, Population Structure, and F-Statistics. Genetics 202:1485–1501.

Pickrell JK, Patterson N, Loh P-R, Lipson M, Berger B, Stoneking M, Pakendorf B, Reich D. 2014. Ancient west Eurasian ancestry in southern and eastern Africa. PNAS 111:2632–2637.

Pickrell JK, Pritchard JK. 2012. Inference of Population Splits and Mixtures from Genome-Wide Allele Frequency Data.Tang H, editor. PLoS Genetics 8:e1002967.

Pickrell JK, Reich D. 2014. Toward a new history and geography of human genes informed by ancient DNA. Trends in Genetics 30:377–389.

Price AL, Tandon A, Patterson N, Barnes KC, Rafaels N, Ruczinski I, Beaty TH, Mathias R, Reich D, Myers S. 2009. Sensitive Detection of Chromosomal Segments of Distinct Ancestry in Admixed Populations. PLOS Genetics 5:e1000519.

Pritchard JK, Stephens M, Donnelly P. 2000. Inference of population structure using multilocus genotype data. Genetics 155:945–959.

Prüfer K, Racimo F, Patterson N, Jay F, Sankararaman S, Sawyer S, Heinze A, Renaud G, Sudmant PH, de Filippo C, et al. 2014. The complete genome sequence of a Neanderthal from the Altai Mountains. Nature 505:43–49.

Qin P, Stoneking M. 2015. Denisovan ancestry in East Eurasian and native American populations. Molecular biology and evolution 32:2665–2674.

Raghavan M, Skoglund P, Graf KE, Metspalu M, Albrechtsen A, Moltke I, Rasmussen S, Stafford TW, Orlando L, Metspalu E, et al. 2014. Upper Palaeolithic Siberian genome reveals dual ancestry of Native Americans. Nature 505:87–91.

Raghavan M, Steinrücken M, Harris K, Schiffels S, Rasmussen S, DeGiorgio M, Albrechtsen A, Valdiosera C, Ávila-Arcos MC, Malaspinas A-S, et al. 2015. Genomic evidence for the Pleistocene and recent population history of Native Americans. Science [Internet] 349. Available from: https://science.sciencemag.org/content/349/6250/aab3884

Rasmussen M, Guo X, Wang Y, Lohmueller KE, Rasmussen S, Albrechtsen A, Skotte L, Lindgreen S, Metspalu M, Jombart T, et al. 2011. An Aboriginal Australian genome reveals separate human dispersals into Asia. Science 334:94–98.

Reich D. 2018. Who we are and how we got here: Ancient DNA and the new science of the human past. Oxford University Press

Reich D, Green RE, Kircher M, Krause J, Patterson N, Durand EY, Viola B, Briggs AW, Stenzel U, Johnson PLF, et al. 2010. Genetic history of an archaic hominin group from Denisova Cave in Siberia. Nature 468:1053–1060.

Reich D, Patterson N, Campbell D, Tandon A, Mazieres S, Ray N, Parra MV, Rojas W, Duque C, Mesa N, et al. 2012. Reconstructing Native American population history. Nature 488:370–374.

Reich D, Thangaraj K, Patterson N, Price AL, Singh L. 2009. Reconstructing Indian population history. Nature [Internet] 461. Available from: https://doi.org/10.1038/nature08365

Sankararaman S, Patterson N, Li H, Pääbo S, Reich D. 2012. The Date of Interbreeding between Neandertals and Modern Humans. PLOS Genetics 8:e1002947.

Sankararaman S, Sridhar S, Kimmel G, Halperin E. 2008. Estimating Local Ancestry in Admixed Populations. The American Journal of Human Genetics 82:290–303.

Schraiber JG, Akey JM. 2015. Methods and models for unravelling human evolutionary history. Nature Reviews Genetics 16:727–740.

Shinde V, Narasimhan VM, Rohland N, Mallick S, Mah M, Lipson M, Nakatsuka N, Adamski N, Broomandkhoshbacht N, Ferry M, et al. 2019. An Ancient Harappan Genome Lacks Ancestry from Steppe Pastoralists or Iranian Farmers. Cell 179:729-735.e10.

Skoglund P, Jakobsson M. 2011. Archaic human ancestry in East Asia. PNAS 108:18301–18306.

Skoglund P, Malmström H, Omrak A, Raghavan M, Valdiosera C, Günther T, Hall P, Tambets K, Parik J, Sjögren K-G, et al. 2014. Genomic diversity and admixture differs for Stone-Age Scandinavian foragers and farmers. Science 344:747–750.

Skoglund P, Mathieson I. 2018. Ancient Genomics of Modern Humans: The First Decade. Annual Review of Genomics and Human Genetics 19:381–404.

Skoglund P, Posth C, Sirak K, Spriggs M, Valentin F, Bedford S, Clark GR, Reepmeyer C, Petchey F, Fernandes D, et al. 2016. Genomic insights into the peopling of the Southwest Pacific. Nature 538:510–513.

Stringer C. 2016. The origin and evolution of Homo sapiens. Philosophical Transactions of the Royal Society B: Biological Sciences 371:20150237.

Tätte K, Pagani L, Pathak AK, Kõks S, Duy BH, Ho XD, Sultana GNN, Sharif MI, Asaduzzaman M, Behar DM, et al. 2019. The genetic legacy of continental scale admixture in Indian Austroasiatic speakers. Scientific Reports 9:3818.

Tishkoff SA, Reed FA, Friedlaender FR, Ehret C, Ranciaro A, Froment A, Hirbo JB, Awomoyi AA, Bodo J-M, Doumbo O, et al. 2009. The Genetic Structure and History of Africans and African Americans. Science 324:1035–1044.

Vernot B, Akey JM. 2015. Complex history of admixture between modern humans and Neandertals. The American Journal of Human Genetics 96:448–453.

Wangkumhang P, Hellenthal G. 2018. Statistical methods for detecting admixture. Curr. Opin. Genet. Dev. 53:121–127.

Yan J, Patterson N, Narasimhan V. 2019. miqoGraph: Fitting admixture graphs using mixed-integer quadratic optimization. bioRxiv:801548.